LEARNING INFLECTIONAL PARADIGMS USING PARALLEL CORPORA

Michelle Dionisio

Supervisors: Philipp Koehn and Mirella Lapata

Master of Science
in
Speech and Language Processing
Linguistics and English Language
School of Philosophy, Psychology and Language Sciences
University of Edinburgh

2006

ABSTRACT

This thesis examines the learning of French inflectional paradigms using parallel corpora. An unsupervised method of accomplishing this is presented, and is contrasted with two other methods which make use only of monolingual corpora. It is found that the combination of all three methods yields the best overall results. The combined model achieves final scores of 54.1% precision, 70.3% recall, and 61.1% f-measure when evaluation against a manually-created gold standard is performed. The use of parallel corpora is found to introduce a performance gain with respect to the learning of irregular morphology.

ACKNOWLEDGEMENTS

I would like to thank Dr. Philipp Koehn and Dr. Mirella Lapata for their kind and patient guidance throughout this project.

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

(Michelle Dionisio)

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Chapter 1

INTRODUCTION

Parallel corpora, which are bilingual sentence-aligned texts, have been shown to be a rich resource for various natural language processing endeavors, such as machine translation and paraphrase generation (Callison-Burch and Bannard (2005)). In this study, the use of parallel corpora in the automatic induction of morphology is examined. By their nature, parallel corpora are an excellent source of information for drawing correspondences between a pair of languages – in addition to being sentence-aligned, these bilingual texts are also word-aligned. Thus, each word in one language is linked to its corresponding counterpart in the other language. Because they are correspondences at the word level, word alignments can be used to draw connections between patterns of word structure in each of the two languages in the parallel corpus. By exploiting these word alignments to project morphological analysis, the bilingual links inherent in the parallel corpus are made use of to automatically learn inflectional morphology.

1.1 Morphology[1]

Morphology is the study of word structure. In morphological theory, words are taken to be composed of meaning-bearing units called *morphemes*. Morphemes can be further classified into *roots* and *affixes*, with *roots* referring to the main morpheme comprising a word and *affixes* referring to those morphemes which when tacked onto the root, signal changes in the meaning of the word.

Morphemes combine in two main ways to form words: via *inflection*, and via *derivation*. In inflection, an affix with a grammatical function (such as plural marking or case marking) combines with a word root, yielding a word which still

[1] The information in this section was mainly drawn from Jurafsky and Martin (2000).

belongs to the same part of speech class as the original root, but which also now carries additional information allowing the word to conform to grammatical requirements such as agreement. In contrast, in derivation, an affix combines with a word stem to form a word of a (usually) different part of speech class. Examples of both inflectional and derivational morphology in English are given below:

- Inflection
  Plural marking: cat (Noun) + -s → cats (Noun)
  Tense marking: stitch (Verb) + -ed → stitched (Verb)

- Derivation
  Nominalization: characterize (Verb) + -ation → characterization (Noun)
  Predicativization: character (Noun) + -ize → characterize (Verb)

This study is limited to the learning of inflectional morphology. Derivational variants of a word are not considered to be part of the paradigm which the methods are expected to learn. To make the distinction more concrete, consider the root word *character*, which has the following inflections and derivations:

- Root: character

- Inflections: characters

- Derivations: characterize, characterizations

The methods presented in this paper deal only with learning the morphology of the inflectional paradigm [character, characters]. The *inflectional paradigm* of a word is taken to be the set of words comprised of the root word itself, plus all its inflected forms. Thus the inflectional paradigm of *character* is the group of words [character, characters].

The derivational variants of *character* (*characterize* and *characterizations*) are not taken to be part of *character*'s inflectional paradigm. Rather, in this study they are treated as root forms in their own right, giving rise to their own respective inflectional paradigms. *Characterize*, for example, has the inflectional paradigm [characterize, characterizes, characterized], while *characterization* has its own inflectional paradigm [characterization, characterizations].

A brief description of some main characteristics of English and French morphology follows. English and French are the two languages investigated in this study.

1.1.1 English morphology

English inflectional morphology is quite simple – in general, nouns and verbs are the only part of speech classes that undergo inflection, and the inventory of inflectional affixes itself is relatively small. Nouns in English can take on two kinds of inflection: plural marking and possessive marking. However, the English data used in this study has been preprocessed to separate the possessive marker *–'s* from the words it is attached to. Thus, the only inflection left intact on nouns in the English data is the plural marker *–s* (and its variants). Plural marking in English is signaled with the use of the canonical suffix *–s* and its variants *–es* and *–en*. The spelling of the root word may undergo changes during this suffixation.

- Examples of plural inflection in English:
  knife + -s → knives
  class + -es → classes
  ox + -en → oxen

English verbal inflection is a bit more complicated. Verbs can be divided into two classes: those that are regular in their conjugation, and those that are

irregular. Regular verbs have predictable inflection patterns which involve the suffixes *–s, –ing,* and *–d/–ed,* as shown below:

Morphological Form Classes	Regularly Inflected Verbs			
stem	walk	merge	try	map
–s form	walk*s*	merge*s*	trie*s*	map*s*
–ing participle	walk*ing*	merg*ing*	try*ing*	mapp*ing*
past form or *–ed* participle	walk*ed*	merg*ed*	trie*d*	mapp*ed*

Table 1. Regular verbal inflection in English[2]

In contrast, irregular verbs do not display such a predictable pattern of inflection:

Morphological Form Classes	Irregularly Inflected Verbs			
stem	eat	catch	cut	be
–s form	eat*s*	catche*s*	cut*s*	am is are
–ing participle	eat*ing*	catch*ing*	cutt*ing*	be*ing*
past form	ate	caught	cut	was were

[2] Jurafsky and Martin (2000), p. 62

–*ed* participle	eaten	caught	cut	been

Table 2. Irregular verbal inflection in English[3]

As explained earlier, derivational morphology in English generally involves nominalization (making nouns from verbs and adjectives) and predicativization (making verbs from nouns and adjectives). In addition, adjectives are also formed from nouns and verbs via derivational morphology. From adjectives, adverbs can be derived through the suffixation of –*ly*. Some common English derivational suffixes are given below:

Process	Suffix	Root Word	Derived Word
Nominalization	–*ation*	computerize (V)	computeriz*ation* (N)
	–*ness*	fuzzy (Adj)	fuzzi*ness* (N)
Predicativization	–*ize*	tender (Adj) digit (N)	tender*ize* (V) digit*ize* (V)
Adjectivization	–*less*	clue (N)	clue*less* (Adj)
Adverbization	–*ly*	clueless (Adj)	clueless*ly* (Adv)

Table 3. Derivational suffixes in English[4]

[3] Adapted from Jurafsky and Martin (2000), p. 63

[4] Adapted from Jurafsky and Martin (2000), p. 64

1.1.2 French morphology[5]

The inflectional morphology of French nouns is similar to that of English nouns in that French nouns also get inflected for plural marking, via the pluralizing suffixes *–s, –x,* and *–aux*:

- Examples of plural inflection in French nouns:
  enfant + -s → enfants
  gateau + -x → gateaux
  cheval + -aux → chevaux

Unlike English, however, French nouns do not get inflected for possessive marking, since possession is signaled via the use of the separate word *de* 'of'.

In addition, French adjectives also get inflected for plural marking:

- Examples of plural inflection in French adjectives:
  intéressant + -s → intéressants
  beau + -x → beaux
  verbal + -aux → verbaux

French adjectives are also inflected for gender agreement. The suffix *–e* is added onto the masculine forms of adjectives to create the feminine forms. This suffixation often brings about changes in orthography:

- Examples of gender agreement inflection in French adjectives:
  intéressant + -e → intéressante
  heureux + -e → heureuse
  ancient + -e → ancienne

[5] The information on French verbal morphology in this section was mainly drawn from Kendris (1996).

French verbal inflection is more complicated than that of English. Verbs in French that conjugate regularly can be divided into three main classes according to the endings of their roots: namely, *–er* verbs, *–ir* verbs, and *–re* verbs. Each of these classes has its own unique pattern of inflection, as shown below:

- Inflectional paradigm for the verb *parler* 'to speak' (*–er* verb):
  [parle, parles, parlons parlez, parlent, parlais, parlait, parlions, parliez, parlaient, parlai, parlas, parla, parlâmes, parlâtes, parlèrent, parlerai, parleras, parlera, parlerons, parlerez, parleront, parlerais, parlerait, parlerions, parleriez, parleraient, parlasse, parlasses, parlât, parlassions, parlassiez, parlassent, parlé, parlant][6]

- Inflectional paradigm for the verb *finir* 'to finish' (*–ir* verb):
  [finis, finit, finissons, finissez, finissent, finissais, finissait, finissions, finissiez, finissaient, finîmes, finîtes, finirent, finirai, finiras, finira, finirons, finiriez, finiraient, finirais, finirait, finirions, finiriez, finiraient, finisse, finisses, finît, fini, finissant][7]

- Inflectional paradigm for the verb *attendre* 'to wait' (*–re* verb):
  [attends, attend, attendons, attendez, attendant, attendais, attendait, attendions, attendiez, attendaient, attendis, attendit, attendîmes, attendîtes, attendirent, attendrai, attendras, attendra, attendrons, attendrez, attendront, attendrais, attendrait, attendrions, attendriez, attendraient, attende, attendes, attendions, attendiez, attendant, attendisse, attendisses, attendît, attendissions, attendissiez, attendissent, attendu, attendant][8]

---

[6] Adapted from Kendris (1996), p. 324

[7] Adapted from Kendris (1996), p. 225

[8] Adapted from Kendris (1996), p. 56

The inflectional paradigms above display only the unique inflected forms that each of the three example verbs can take on – they are not meant to be an illustration of the full conjugation of these verbs with respect to person, number, and tense. As can be seen from the three example paradigms above, the different verb classes in French have varying numbers of inflectional forms. The *–ir* verb class, for example, has a lesser number of unique inflectional forms than the *–re* verb class.

There are several irregularly-conjugating verbs in French. Many of these verbs have the same root word endings (*–er, –ir,* or *–re*) as regularly-conjugating verbs, yet their patterns of inflection are irregular. An example of such an irregular verb is *avoir* 'to have', which, although possessing an *–ir* ending, has a completely irregular pattern of inflection:

- Inflectional paradigm for the irregular verb *avoir*:
  [avoir, ai, as, a, avons, avez, ont, avais, avait, avions, aviez, avaient, eus, eut, eûmes, eûtes, eurent, aurai, auras, aura, aurons, aurez, auront, aurais, aurait, aurions, auriez, auraient, aie, aies, ait, ayons, ayez, aient, eusse, eusses, eût, eussions, eussiez, eussent, eu, ayant][9]

As in English, derivational morphology in French involves the formation of nouns, verbs, adjectives and adverbs. Some examples of French derivational suffixes are shown below:

Process	Suffix	Root Word	Derived Word
Nominalization	*–ation*	caractériser (V)	caractéris*ation* (N)

[9] Adapted from Kendris (1996), p. 61

	–ité	rapide (Adj)	rapid*ité* (N)
Predicativization	*–ifier*	simple (Adj)	simpl*ifier* (V)
Adjectivization	*–aire*	planète (N)	planet*aire* (Adj)
Adverbization	*–ment*	rapide (Adj)	rapidement (Adv)

Table 4. Derivational suffixes in French[10]

As explained earlier, in this study derivational variants of a given root word are treated as root words in their own right. Thus the verb *caractériser* 'to characterize' has its own inflectional paradigm (consisting of *caractériser* plus all its inflected forms), while its derived noun *caractérisation* 'characterization' will have its own, separate inflectional paradigm (namely, [caractérisation, caractérisations]).

1.2 Previous Work

The goal of this study is to investigate unsupervised approaches to learning inflectional paradigms using parallel corpora. Thus the following survey of the literature is limited to approaches which make use of unsupervised or minimally supervised morphology induction methods. This previous work can be divided into two categories: those utilizing monolingual corpora, and those utilizing parallel corpora.

1.2.1 Approaches using monolingual corpora

Yarowsky and Wicentowski (2000)

Yarowsky and Wicentowski (2000) present an algorithm for inducing inflectional morphology from a monolingual corpus with no direct supervision. The algorithm specifies the combination of four models for aligning an inflection with

[10] Adapted from Lessard (1996), ch. 5

its root. These models are based on relative corpus frequency, contextual similarity, weighted Levenshtein distance, and incrementally retrained stem change probabilities. Notably, Yarowsky and Wicentowski's approach handles the learning of both regular and irregular inflectional morphology.

Inflectional morphological analysis is treated as essentially an *alignment* task – a probabilistic alignment between inflections and roots is first estimated. These collected alignments can then actually be used as a morphological analyzer via simple lookup – for each given inflection, we can simply look up its alignment to find its root. However, Yarowsky and Wicentowski go a step further and use a weighted subset of these aligned <inflection, root> pairs to train a supervised morphological analysis learner. This trained learner can then be used as a stand-alone morphological analyzer or as a probabilistic scoring component to iteratively improve the alignments of inflections and roots.

The resources required for this approach are as follows:

- A table of the parts of speech of the relevant language and the canonical suffixes for each part of speech

- A large unannotated text corpus

- A list of candidate roots, along with a way of identifying parts of speech of the remaining vocabulary using context or tag sequence, *not* morphological analysis. The part of speech tag estimates function to limit unrestricted word-to-word alignments over the entire vocabulary.

- A list of consonants and vowels of the language

The following resources are optional:

- A list of common function words – used in extracting context similarity features

- Distance or similarity tables previously generated by this algorithm for other languages – used as seed information

The first method Yarowsky and Wicentowski present for aligning inflections and roots is based on frequency similarity. This method operates on the insight that an inflection and its root should have close relative frequencies. And this does appear to be borne out: *sang* and *sing*, for example, have a relative frequency ratio of 1427/1204 (or 1.19/1), while the morphologically unrelated words *singed* and *sing* have a ratio of 9/1204 (or 0.007/1), which is quite dissimilar.

However, since some inflections are rarer than their root forms and can thus be expected to occur less frequently than their roots, expected frequency distributions are calculated to properly rank the inflection/root ratios based on how well they fit or deviate from the expected frequencies.

Another alignment model presented in Yarowsky and Wicentowski's paper is based on context similarity. Cosine similarity is computed between vectors of context features, with the result that inflectional variants of the same word receive high similarity scores, since they have very similar argument distributions and selectional preferences.

The third alignment model presented in the paper makes use of weighted Levenshtein distance. By calculating the overall stem edit distance, related inflectional variants of a given stem can be found. Instead of giving all string edits equal cost, though, this weighted version penalizes changes involving consonants more than those changes which involve only vowels. The rationale behind this is that in the morphology of most languages, consonants have a lower

probability of changing during inflection than vowels do. Thus, if a stem edit seen in a given inflected word involves a consonant, it is less likely that this inflected word is a morphologically related variant of the original stem. The costs assigned by the weighted Levenshtein distance measure reflect this expectation.

Lastly, Yarowsky and Wicentowski present an alignment model based on morphological transformation probabilities. The probability of an inflection given a root and a suffix is calculated via the probability of the stem change involved given the root and the suffix. The scores from each of the four models are then scaled and combined to give one score per candidate root for a given inflection.

Evaluation focused on the morphological analysis of English past tense verbs. Although the models on their own did not fare very well (the frequency similarity, Levenshtein distance, and context similarity models yielded accuracy scores of 10%, 31%, and 28%, respectively), the full combined model performed impressively well, yielding a 99.2% accuracy score on the test set.

Schone and Jurafsky (2001)

Schone and Jurafsky (2001) is another knowledge-free approach to the induction of inflectional morphology. Cues from orthography, semantics, and syntactic distributions are used to induce morphological relationships. The algorithm takes a large corpus as its input and outputs sets of morphologically related words. These conflation sets are made up of the inflected and derived variants of a given word – there is no distinction made between inflectional morphology and derivational morphology. All inflectional and derivational variants of a given word are considered to be part of its conflation set.

First, a list of pairs of potentially morphologically related words is generated from an untagged corpus. This is done by identifying word pairs which differ only by a

prefix, or by a suffix, or by a circumfix. Semantic vectors are then determined for each word. These semantic vectors are correlated by calculating normalized cosine scores, and those word pairs whose normalized cosine scores are most likely to be non-random (defined via a probability threshold) are accepted as displaying a valid relationship of morphological relatedness. By grouping together those words that have been determined to be morphologically related in this way (i.e. via their normalized cosine scores), conflation sets of related words can be built.

Next, minimum edit distance is used to calculate an orthography-based probability, which is then combined with the semantic probability obtained earlier via latent semantic analysis. In addition, a syntax-based probability is also determined by identifying word sets which occur around a given word. The normalized cosine score and corresponding probability of these contextual words are then calculated for each potentially morphologically related word pair, and those word pairs with resulting probabilities exceeding a predefined threshold are then deemed to be validly related.

Lastly, valid morphological variants that have not been captured by the combined semantic, orthographic, and syntactic probabilities are identified through transitive closure. For example, if word *X* is found to be related to word *Z*, and word *Y* is also found to be related to word *Z,* then it can be surmised that words *X* and *Y* are related as well.

Evaluation was done by comparing the conflation sets of morphologically related words outputted by the algorithm against their corresponding sets in the CELEX lexicon (Baayen *et al.* (1993)). When scoring for suffixing, an f-score of 88.1% was obtained; for circumfixing, it was 84.5%. The successive additions of each of the probability measures described earlier resulted in improved performance –

from 85.2% to 88.1% in the case of suffixing, and from 82.2% to 84.5% for circumfixing.

1.2.2 Approaches using parallel corpora

Rogati et al. (2003)

Rogati *et al.* (2003) make use of a parallel corpus to build an unsupervised Arabic stemmer. The only resources required by their method are an English stemmer and a small parallel (Arabic-English) corpus. The English stemmer is used to stem the English side of the corpus, while the Arabic side of the corpus is stemmed using an initial guess (i.e. either random stemming or using a simple language-specific rule). Once both sides of the parallel corpus are stemmed, a translation model is built which establishes correspondences between stems in the Arabic and stems in the English.

The translation model is a matrix of translation probabilities *p(Arabic stem|English stem)*, which are refined iteratively using the EM algorithm. The Arabic side of the corpus is re-stemmed using scores for the stems calculated by taking into account the stem's translation probability, and the conditional probabilities of its prefix and suffix. A more accurate translation model is then built and the process is repeated.

The performance of the unsupervised stemmer was evaluated by examining its agreement with a proprietary, rule-based Arabic stemmer. Agreement was 87.5%. A second evaluation was performed, this time task-based Arabic information retrieval. For this evaluation, it was found that the unsupervised stemmer performs at 93.96% agreement with the proprietary stemmer.

Yarowsky et al. (2001)

Yarowsky *et al.* (2001) describe a method for automatically inducing a morphological analyzer by making use of bilingual parallel corpora and an

existing lemmatizer. The morphological analysis of one of the languages in the parallel corpus is projected onto the other language via word alignment probabilities.

A given inflection is associated with its correct root by making use of the word alignments the inflection and root have in common as a bridge. Yarowsky *et al.* used a French-English parallel corpus, and associated French verbal inflections with their correct roots by first finding which English words the inflections were aligned with, and then seeing which other French words were aligned to these English words. However, since direct associations (links) between inflections and their roots are rare (since inflected forms in the French tend to be aligned to inflected forms in the English, while root forms tend to only be aligned to root forms), it is necessary and advantageous to make use of an existing lemmatizer to lemmatize the English side (or whatever the bridge language will be). In this way, inflections in the French which are aligned to inflections in the English will now be aligned to roots in the English (since all inflected forms in the English will have been transformed by the lemmatizer into roots), and thus there is now a direct link between the inflected French forms and the root French forms, via the English stems.

This bridging is formalized as the following similarity measure:

$$P_{mp}(F_{root} \mid F_{infl}) = \Sigma_i P_a(F_{root} \mid E_{lem\ i})\ P_a(E_{lem\ i} \mid F_{infl})$$

where:

- P_{mp} refers to the morphology projection probability we are calculating with this formula. It tells us how probable a certain candidate French root is for a given inflected French word, based on the similarity calculation expressed in the formula.

- P_a refers to alignment probabilities.

- F_{root} refers to candidate French roots.

- F_{infl} refers to inflected French words.

- E_{lem} refers to English lemmas with which both the candidate French root and the inflected French word are aligned (i.e., the aligned English lemmas they have in common.). The index *i* indicates that the product of the alignment probabilities are summed over all such English lemmas which both the candidate French root and inflected French word share an alignment with. In other words, this calculation is performed for every English lemma that is involved in alignments with both the candidate French root and the inflected French word.

Evaluation is done by checking to see if the system has selected the correct root for a given inflection. Yarowsky *et al.* report a precision score of 99.2% for 77.9% of types (inflected French verbs), which constitutes 99.4% of the tokens in their corpus. These scores are for a corpus of 12 million words. For the smallest corpus they used (120,000 words), precision was 96.2% for 0.095% of types (90.1% of tokens). Coverage was found to increase along with the size of the corpus.

The morphology projection method was then augmented with a trie-based morphology model for improved performance. Further performance boosts were achieved with the use of multiple parallel translations (namely, different versions of the Bible for one language) and multiple bridge languages for the morphology projection. With these augmentations, precision scores rose to 99.4%, with full coverage.

1.3 Motivation for this study

Morphological analysis is a useful preliminary step in many NLP tasks. Parsing text, building machine-readable dictionaries, doing machine translation – all these endeavors benefit from the addition of morphological information. Morphology gives clues to linguistic structure beyond the word level – for example, it is a good cue to case, part of speech, number, person, and gender in many languages. This makes morphological analysis a useful step for extracting features from words, which could then be used by machine learning methods for parsing or machine translation.

However, it is both expensive and time- and labor-intensive to morphologically annotate corpora by hand or build rule-based morphological analyzers. Thus it is desirable to handle morphological analysis using as unsupervised an approach as possible. In this study, the problem of automatic morphology induction is tackled from such a perspective. An unsupervised method for the induction of inflectional French morphology using parallel corpora is presented. There are relatively few past works that have dealt with morphology induction using parallel corpora. The goal of this study is to investigate the possibilities offered by this commonly available resource with respect to the learning of inflectional morphology, and to see if any benefits can be derived from its use.

Chapter 2

METHODOLOGY

As described earlier, the main goal of this study is to investigate the use of parallel corpora in learning inflectional French morphology. Towards this end, three separate unsupervised morphology induction methods were implemented: one using parallel corpora, and two using monolingual corpora. Evaluation was performed by comparing the methods' output against a manually created gold standard. The following sections describe the nature of the resources used, the creation of the gold standard, and the implementation of the methods.

2.1 Resources Used

For the approach implemented in this study, the following resources are required: a parallel (French-English) corpus, an English lemmatizer, and a dictionary of French inflectional paradigms. Each of these resources is described in more detail below.

2.1.1 Corpus

The corpus used was the French-English section of the Europarl corpus (Koehn (2002), (2005)). This is a bilingual parallel text comprised of collected proceedings of the European Parliament. Each sentence on the French side of the corpus is aligned with a sentence on the English side; furthermore, each word in the French is aligned with a word (or words) in the English. The data has been preprocessed to remove extraneous annotations such as speaker tags, and has been tokenized as well, such that punctuation marks are a separate token by themselves. A lowercased version of the corpus was used in this study to ensure that the orthography-based learners would not make erroneous distinctions between the capitalized version of a word and its lowercased form.

The corpus was obtained in the form of three files: one file containing the French text, another containing the English text, and a third file containing the alignments. To better illustrate the nature of the data, here is the first sentence of the French text:

```
je déclare reprise la session du parlement
européen , qui avait été interrompue le jeudi 28
mars 1996 .
```

Here is the first sentence of the English text, with which the above French sentence is aligned:

```
i declare resumed the session of the european
parliament adjourned on thursday , 28 march 1996 .
```

In both the French and the English text files, there is one sentence per line.

The alignments were contained in their own separate file. Here is the first line of the alignments file, which corresponds to the first line (and therefore, the first sentence) in both the French and the English texts:

```
0-0 1-1 2-2 3-3 4-4 5-5 5-6 6-8 7-7 7-8 8-12 9-9
10-9 11-9 12-9 13-10 14-11 15-13 16-14 17-15 18-16
```

As described in Koehn (2002) and (2005), Europarl parallel corpora are automatically sentence-aligned. This means that each sentence in the French is aligned with its counterpart sentence in the English. Each aligned sentence pair shares the same line number between the two text files – for example, line number 1 in the French text file contains the first French sentence, which is aligned with the first English sentence, which itself is on line number 1 in the English text file. Thus, line number 1 in the French text and line number 1 in the English text are aligned with each other. This is true for all lines in the parallel corpus – each line of text in the first language is aligned with its counterpart line in the second language.

This correspondence extends to the alignment file. Each line in the alignments file contains the word alignments for the corresponding sentences (lines) in the French and English texts. For example, line number 1 in the alignments file (shown earlier) displays the word alignments for the aligned sentence pair of line number 1 in the French and line number 1 in the English. In essence, the sentence alignments link the three separate files of the corpus together (the French text, the English text, and the alignments file). Because of these sentence alignments, it is assured that line *x* in the French text corresponds to line *x* in the English text, and that furthermore, both these lines correspond to line *x* in the alignments file. This is borne out by the fact that all three corpus files have an identical number of lines (688,031 lines in each file).

Word alignments are represented in the alignments file via a pair of numbers separated by a hyphen (for example, *0-0*). This representation assumes that words in a sentence are numbered starting from 0. The direction of the alignment (which language is represented by the first number and which by the second) depends upon the parallel corpus used; for this study, it is French → English. Thus, the word alignment *0-0* indicates that the first French word (word 0) is aligned to the first English word (word 0) in a given pair of aligned French and English sentences.

The table below displays some of the word alignments for the first aligned sentence pair in the parallel corpus. Each pair of aligned words is in its own box. The number under each word indicates the numbering of that word within the sentence it belongs to.

je	déclare	reprise	la	session	du	du	parlement	européen
0	1	2	3	4	5	5	6	7

i	declare	resumed	the	session	of	the	parliament	european
0	1	2	3	4	5	6	8	7

Table 5. Word alignments for first sentence pair

Note that the French word *du* (index number 5) appears in two word alignments: once with the English word *of,* and once with the English word *the.* This is appropriate since *du* does indeed translate to *of the* in English – it is a collapsing of the French word *de* 'of', and *le* 'the'.

The word alignments were automatically generated using GIZA++ (Och and Ney (2003)). As such, the alignments are not perfect – there are erroneously aligned word pairs.

For almost all of the experiments in this study, a subset of the parallel corpus was used (specifically, 5164 lines of French text, 5164 lines of English text; approximately 120,000 words per language) instead of the corpus in its entirety. This was done partly to keep processing time at a reasonable duration during system development. More importantly, Yarowsky *et al.* (2001) used a corpus of size 120,000 words in developing and evaluating their morphology projection method. Since their method is also implemented and investigated in this study, it was decided that it would be best to use a corpus of the same size for maximum comparability of results.

2.1.2 English lemmatizer

As discussed in section 2.3.3, for one of the morphology learning methods a lemmatizer was applied to the English side of the parallel corpus in order to collapse English inflections into their root forms. The tool used in this study was the English morphological analyzer *morpha* (Minnen *et al.* (2001)). Given English text, *morpha* returns for each word its lemma and inflectional suffix (if any). Both POS-tagged and untagged English data can be accepted by the lemmatizer as

input; however, Minnen *et al.* recommend that tagged English data be used for maximum lemmatizing accuracy. The example below illustrates the morphological analysis of a sentence in the English data, as performed by *morpha*:

- Original sentence:

```
i declare resumed the session of the european
parliament adjourned on thursday , 28 march 1996 .
```

- After morphological analysis by *morpha:*

```
i declare resume+ed the session of the european
parliament adjourn+ed on thursday , 28 march 1996
.
```

As the above shows, *morpha* returns the root forms of inflected words, along with their inflectional suffixes. Thus, the inflected word *resumed* is broken down into its lemma *resume,* and the suffix *–ed*.

2.1.3 French morphological dictionary

Since this study focused on the learning of French inflectional paradigms, a French morphological dictionary was consulted in order to ensure the completeness and correctness of the paradigms that comprise the gold standard.[11] *Morphalou* is a lexicon of inflected French words which can be freely interrogated on the Internet (Romary *et al.* (2004)). In addition to returning the lemma for a given inflected French word, *morphalou* is also capable of outputting the whole inflectional paradigm for a given French lemma. It is this second function which was exploited during the creation of the gold standard for this study.

One thing that is important to note about *morphalou* is that it inherently maintains a distinction between inflectional and derivational variants of a word. Given a lemma, *morphalou* returns only those words which are inflectional variants of that

[11] The creation of the gold standard is described in section 2.2.

lemma. It does not return derivational variants. For example, if *morphalou* is given the lemma *monde* 'world' as input, it will return the set of words [monde, mondes], which indeed contain only the inflected forms of *monde*. The derived forms of *monde,* such as the word *mondial* 'global', is not included as part of the outputted paradigm.

2.2 Creation of the gold standard

Early on in the study, it was decided that evaluation would be performed by comparing the output of the morphology learning methods against a manually-created gold standard. This approach is similar to that used by Schone and Jurafsky (2001), who compared the conflation sets outputted by their system against corresponding word sets in CELEX. Since the evaluation strategy has a direct bearing on the implementation of the morphology learning methods, the creation and finalization of the gold standard was given first priority.

The gold standard in its final form is made up of word pairs (1,619 in all) built from 100 collected unique inflectional paradigms. To better illustrate the nature of the gold standard, the steps undertaken in creating it are described in detail below.

1. Randomly selected 200 unique words from the French side of the parallel corpus. These 200 words were selected from the 120,000-word monolingual French subset of the parallel corpus used in most of the experiments in this study.[12] This was done using a Python script.

   example: Sample script output:

```
honnêtement, constituée, modalités, wulf-mathies,
incertaines, turin, start, avoir, lorsqu, a4-0101,
```

[12] Section 2.1.1 explains the justification for using a 120,000-word subset of the corpus instead of the whole corpus.

```
européen, play, désastre, naturellement, messages,
déclencheraient, dangers, devons, devrait, …
```

2. From these 200 words, manually filtered out named entities, punctuation, numbers, and non-French words.

   example: After removing named entities `wulf-mathies` and `turin`, as well as the number `a4-0101` and the English words `start` and `play`, the remaining words look like:

```
honnêtement, constituée, modalités, incertaines,
avoir, lorsqu, européen, désastre, naturellement,
messages, déclencheraient, dangers, devons,
devrait, …
```

   After doing the above filtering, the remaining words were further manually inspected to weed out redundant morphological variants. For example, the words *devons* and *devrait* are present among the remaining words in the text box above. *Devons* and *devrait* are morphologically related: they are inflected forms of the verb *devoir* 'must'. Now, the point of collecting these words from the corpus is to use them to build up distinct inflectional paradigms that will make up the gold standard. Having morphologically related words in the pool of gold standard words would lead to the creation and inclusion of redundant paradigms. Therefore, it is best to eliminate morphological variants from the pool of gold standard words at this stage.

   The way this was done was as follows: the list of remaining words was scrutinized from top to bottom. If a given word happened to be a morphological variant of a word that already occurred earlier in the list, this word (the word currently being inspected) was removed from the list. Thus, in the example above, *devrait* would be removed from the list since a related morphological variant, *devons,* already exists in the list.

3. Again using a Python script, randomly selected 100 words from the remaining words. Manually checked words to see if a variety of part of speech classes (noun, verb, adjective, adverb, function words[13]) were represented among these 100 words. Re-ran script to re-generate another 100 randomly selected words and checked words again, repeating as necessary until part of speech variety was adequately achieved.

   example: Sample script output:

```
assisté, modalités, attendre, incertaines, avoir,
désastre, dangers, honnêtement, lorsqu, …
```

4. Obtained the complete inflectional paradigm for each of these 100 words. This was the most time-consuming step in the creation of the gold standard, as it was mostly manually done. The substeps involved were as follows:

   For each of the 100 words:

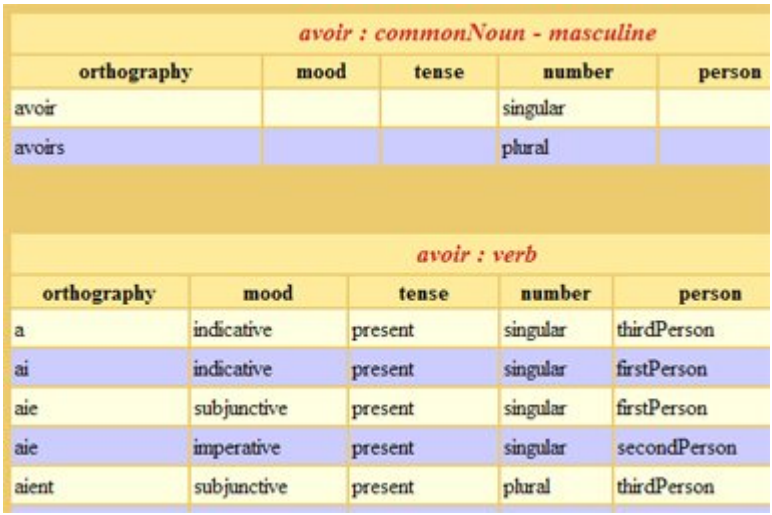
   i) Typed the word into the lemma field (the first search box) at the *morphalou* search page[14] to see if it has any inflected forms.

      If the word is not a lemma but is rather an inflected form (as many of the words in the gold standard are), *morphalou* returns the error message 'No lexical entry found'. In this case, the next action taken was step (iii) below.

[13] Function words in French (such as prepositions, determiners, complementizers, and question words) can have different inflected forms. Some examples are [*de, du*] 'of', [*le, la, les, l'*] 'the', [*lorsque, lorsqu'*] 'when', [*que, qu'*] 'what', and [*quel, quelle, quels, quelles*] 'which', among others.

[14] http://actarus.atilf.fr/morphalou/morphalou_req.html

If the word is a lemma, *morphalou* outputs its full inflectional paradigm. Note that it is possible for a given surface form to have more than one inflectional paradigm. An example of this is the word *avoir*, which can be either a verb meaning 'to have', or a noun meaning 'asset'. When *avoir* is typed into the lemma field at the *morphalou* search page, the following result is returned:



avoir : commonNoun - masculine

orthography	mood	tense	number	person
avoir			singular	
avoirs			plural	

avoir : verb

orthography	mood	tense	number	person
a	indicative	present	singular	thirdPerson
ai	indicative	present	singular	firstPerson
aie	subjunctive	present	singular	firstPerson
aie	imperative	present	singular	secondPerson
aient	subjunctive	present	plural	thirdPerson

Figure 1. *Morphalou* result page showing inflectional paradigms for *avoir*

As the above figure shows, there are two inflectional paradigms for the surface form *avoir:* [avoir, avoirs] – namely, the singular and plural forms of the noun *avoir* meaning 'asset'; and [a, ai, aie, aient, …] – namely, the different inflected forms of the verb *avoir* meaning 'to have'.[15]

ii) This outputted inflectional paradigm was then saved to file. This was done by selecting all the inflected forms in the table, then copying and pasting them into a text file.

For those words (such as *avoir*) with more than one inflectional paradigm, the separate inflectional paradigms returned by *morphalou* were collapsed together into one. Thus, *avoir* was considered as having the single inflectional paradigm [avoir, avoirs, a, ai, aie, aient, …]. The motivation for doing this is the fact that the words in the gold standard are being treated as isolated surface forms – they were randomly plucked out of the corpus, and thus it is impossible to tell what their original meanings were in the text. Because of this, none of the possible inflectional paradigms for a given surface form can be ruled out, and so all of them are simply combined into one paradigm.

The text file of saved inflectional paradigms was formatted such that there was one paradigm per line. Each paradigm was comprised of a succession of words separated by whitespace.

example: Sample lines in text file of inflectional paradigms:

```
a ai aie aient aies ait as aura aurai  …
danger dangers
honnêtement
modalité modalités
attend attendaient attendais attendait …
lorsqu lorsque
```

iii) Those words in the gold standard which are inflected forms, for which *morphalou* yielded the error message 'No lexical entry found' when they were typed into the lemma search box in step (i), were then typed into the inflection search box on the *morphalou* search page. This is the second search box on the page. For each word, *morphalou* returned the lemma(s) to which that word belonged.

---

[15] The full inflectional paradigm for the verb *avoir* is given in Section 1.1.2.

For example, when the inflected form *assisté* was typed into the inflection search box, *morphalou* returned the following result:

assisté	
lemma	**category**
assister	verb
assisté	commonNoun
assisté	adjective

Figure 2. *Morphalou* result page showing lemmas for *assisté*

According to *morphalou,* the inflected form *assisté* can belong to any of three different paradigms: that of the verb *assister* 'to assist', that of the noun *assisté* 'assisted', and that of the adjective *assisté* 'assisted'. The next step was then to enter each of these lemmas into the lemma search box in order to obtain their inflectional paradigms. The inflectional paradigms obtained for each of these three lemmas were as follows:

- Inflectional paradigm for *assisté* (noun) – consists of the singular and plural forms of the noun:

```
assisté assistés
```

- Inflectional paradigm for *assisté* (adjective) – consists of the singular and plural version of both masculine and feminine forms of the adjective:

```
assisté assistés assistée assistées
```

- Inflectional paradigm for *assisté* (verb): consists of all the inflected forms of the verb:

```
assista assistai assistaient assistais
assistait assistant assistas assistasse
assistassent assistasses assistassiez
assistassions assiste assistent assister
assistera assisterai assisteraient
assisterais assisterait assisteras
assisterez assisteriez assisterions
assisterons assisteront assistes assistez
assistiez assistions assistâmes assistât
assistâtes assistons assistèrent assistés
assisté assistée assistées
```

These three separate paradigms were then collapsed together to form one inflectional paradigm for the gold standard word *assisté*. Note that the words comprising the inflectional paradigms for the noun and adjective lemmas *assisté* (namely, *assisté, assistés, assistée,* and *assistées*) are already part of the inflectional paradigm for the verb *assisté* anyway (words in boldface above). This is because the surface forms of the noun and adjective inflections for *assisté* happen to be identical to the surface form of the past participle for the verb *assister,* which is *assisté* 'assisted'. Like the nouns and adjectives, this past participle gets inflected for number and gender – thus it has variants *assistés, assistée,* and *assistées.* The final paradigm for *assisté* therefore ends up being identical to the inflectional paradigm for the verb shown above.

This sharing of inflected forms was never taken for granted, however, as there are some instances where the noun or adjective inflectional paradigms contain an inflected form that is not present in the verbal inflectional paradigm. Thus, the possible lemmas for each inflected gold standard word were each

methodically looked up, and the inflectional paradigms for these lemmas collected and collapsed together to ensure that all possible inflected forms would be included in the final paradigm for the gold standard word.

5. When all the complete inflectional paradigms for each of the 100 gold standard words had been finalized, a Python script was used to filter out those words within these paradigms which do not occur in the corpus.[16] This step is necessary since the complete inflectional paradigm for a given gold standard word contains inflected forms which are not attested in the corpus (for example, rare verb conjugations). These unattested inflected forms must be removed from the collected paradigms in order for these paradigms to be validly used in evaluating the output of the morphology learning methods. (The morphology learners will only be able to output words that are attested in the corpus, so the gold standard cannot contain words which the learners will never see.)

6. Lastly, another Python script was used to create word pairs out of words belonging to the same paradigm. For example, consider the following paradigm:

```
interrompt interrompue interrompe
```

This is what is left of the complete inflectional paradigm built up for the gold standard word *interrompue* 'interrupted', after those inflected forms unattested in the corpus were weeded out. In making word pairs, the script first alphabetizes the words within each paradigm:

```
interrompe interrompt interrompue
```

[16] The relevant corpus here is still the 120,000-word French monolingual subset of the parallel corpus, described earlier in section 2.1.1.

Then all possible word pairs are built from these words:

```
[interrompe, interrompt]
[interrompe, interrompue]
[interrompt, interrompue]
```

The gold standard in its final form is made up of word pairs like those in the text box above. To better illustrate its appearance, here is a snippet of the gold standard text file. There is one word pair per line in this file:

```
d\xe9crire d\xe9crit
d\xe9crire d\xe9crites
d\xe9crit d\xe9crites
interdit interdite
interdit interdites
interdite interdites
modalit\xe9 modalit\xe9s
doigt
constern\xe9
affirmative
presse pressions
presse press\xe9
```

Figure 3. Snippet of gold standard text file

Notice that there are single words in the gold standard, such as *doigt*, *consterné,* and *affirmative* in the figure above. These single words arose from paradigms that ended up with only one word in them after all inflected forms unattested in the corpus were weeded out. For example, the complete inflectional paradigm for the word *doigt* 'finger' was [doigt, doigts]. However, since *doigts* was unattested in the corpus, it was filtered out, leaving the paradigm with one member: [doigt].

In its final form, the gold standard is comprised of 1,619 word pairs. These 1,619 word pairs contain 391 unique French words (types).

2.3 Morphology learning methods

With the evaluation strategy thus finalized and firmly in place, the next stage in the study was to implement the morphology learning techniques. Since

evaluation was to be done by comparing conflation sets of morphologically related words (*à la* Schone and Jurafsky (2001)), the basic guiding principle behind the implementation stage was to develop a morphology learning method that would be able to group morphologically related words within a text into sets. Thus the output of the method would be sets of related words (in other words, inflectional paradigms) that would then be able to be broken up into word pairs and directly compared against the gold standard.

The three morphology learning methods developed in this study are described below. All three methods were implemented in Python. As previously stated, the goal of these methods is to learn French inflectional morphology. The first two methods, Prefix Similarity and Levenshtein Distance, make use of only monolingual corpora. These two methods provide baseline results against which the performance of the third method (the Projection Method, which uses parallel corpora) is gauged.

2.3.1 Prefix Similarity

The first method implemented used prefix similarity to determine morphological relatedness. In essence, this method checked to see if two given words from the corpus had the same prefix. If they did, then the method considered them as morphologically related and placed them in the same conflation set.

This method required only monolingual data – it made use only of the French side of the parallel corpus. It directly compared the French words to each other, using the following rules to determine prefix similarity:

For each word x in the monolingual French corpus:

> Compare x against each word y in this same corpus and determine their prefix similarity as follows:

1. If *x* and *y* are both either one letter in length, or two letters in length:

   If the first letters of both words are the same, then *x* and *y* are morphologically related.

   example: a is morphologically related to a
   le is morphologically related to la

2. If *x* and *y* are both three letters in length:

   If the first two letters of both words are the same, then *x* and *y* are morphologically related.

   example: dis is morphologically related to dit

3. If *x* and *y* are both four or more letters in length:

   If the first four letters of both words are the same, then *x* and *y* are morphologically related.

   example: dira is morphologically related to dirais

As discussed in the evaluation of this method,[17] other rules and rule combinations were tried out in experiments. The rules described above are those that were found to yield the best results.

Words that are found to be morphologically related are placed into the same conflation set. Thus, the output of this method consists of sets of morphologically related words. These sets of morphologically related words are

[17] Section 3.1 describes the evaluation of the Prefix Similarity method.

the inflectional paradigms determined by the prefix similarity model. A sample of the outputted paradigms (in text format) is shown below:

```
['europ\xe9en', 'europ\xe9enne', 'europ\xe9ennes', 'europ\xe9ens']
['prosp\xe8re', 'prosp\xe8res', 'prosp\xe9rer']
['fond', 'fonde', 'fondent', 'fonder', 'fonds', 'fond\xe9', 'fond\x
['dira', 'dirai', 'dirais', 'dirait']
['dis', 'dit']
['sup\xe9rieur', 'sup\xe9rieure', 'sup\xe9rieures']
['attend', 'attendant', 'attende', 'attendent', 'attendez', 'attend
['vers\xe9', 'vers\xe9es', 'vers\xe9s']
['parlais', 'parlait', 'parle', 'parler', 'parlerai', 'parlez', 'pa
['m\xe9ridionales', 'm\xe9rite', 'm\xe9ritent', 'm\xe9riterait', 'm
['crucial', 'cruciale']
['honneur', 'honn\xeatement']
['b\xe9n\xe9fique', 'b\xe9n\xe9fiques']
['assister', 'assistons', 'assist\xe9']
['quel', 'quelle', 'quelles', 'quels']
```

Figure 4. Output of Prefix Similarity method

2.3.2 Levenshtein Distance

The second morphology learning method implemented made use of Levenshtein distance as a similarity measure. Levenshtein distance is the edit distance between two strings, where edits can be insertions, deletions, or substitutions. These edits are assigned costs (which can be weighted, as in Yarowsky and Wicentowski (2000)), and these costs are then used to calculate the overall edit distance between two strings.

Because the morphology learning methods which make use of monolingual corpora (namely, the Prefix Similarity method and the Levenshtein Distance method) were used in this study to obtain baseline results, the implementation of the Levenshtein Distance method was kept as simple as possible. Thus, all edits – whether insertions, deletions, or substitutions – were uniformly assigned a cost of 1. Given this configuration, the Levenshtein distance between, for example, the words *quel* and *quels* would therefore be 1, since that is the cost incurred by inserting *s* onto *quel* to form *quels*. Similarly, the distance between *le* and *la* would be 1, which is the cost of substituting the *e* in *le* with an *a* to form *la*.

Like the Prefix Similarity method described in the preceding section, the Levenshtein Distance method required only monolingual data. It performed its calculations in the same systematic manner: each word *x* in the monolingual French corpus was considered against every word *y* in this same corpus in turn. For each pair of words *x* and *y* under consideration, the Levenshtein distance between these two words was calculated. If the Levenshtein distance between *x* and *y* was found to be less than or equal to a specified cutoff value, then *x* and *y* were determined to be morphologically related.

As discussed in the evaluation of this method,[18] several different cutoff values were tested in experiments. A cutoff value of 2 was found to yield the best results.

Again like the Prefix Similarity method, the Levenshtein Distance method places words it has deemed to be morphologically related into the same conflation set. Thus the output of the method consists of sets of related words, which can be considered as the inflectional paradigms determined by the Levenshtein distance model. Below is a sample of the outputted paradigms (in text format):

[18] Section 3.2 describes the evaluation of the Levenshtein Distance method.

Figure 5. Output of Levenshtein Distance method

2.3.3 Morphology Projection

The third morphology learning method implemented was the Projection method. Unlike the Prefix Similarity and Levenshtein Distance methods, which make use of only monolingual corpora, the Projection method makes use of parallel corpora. The implementation of this method was based upon the morphology projection similarity measure described in Yarowsky *et al.* (2001).

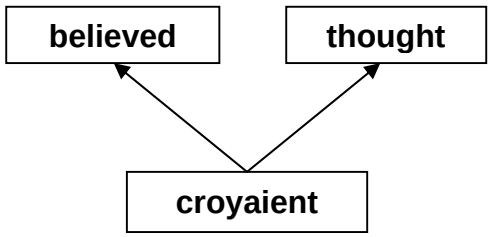
The Projection method exploits the unique feature that parallel corpora introduce to the scene: word alignments. The availability of word alignments as a resource is mainly what differentiates the use of monolingual corpora from the use of parallel corpora. As was discussed earlier in section 1.2.2, Yarowsky *et al.* (2001) use word alignments to project the morphological analysis of one of the languages in the parallel corpus onto the other language in the corpus. Thus the word alignments are used as a bridge over which morphological analysis is projected from one side of the parallel corpus onto the other.

This process can be illustrated as follows. The example below, adapted from Yarowsky *et al.* (2001), involves morphology projection from French to English.

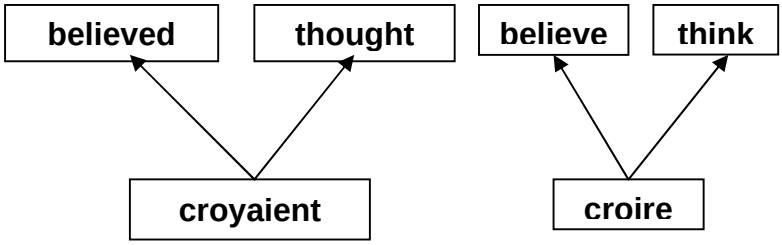
Example:

Given an inflected French word *croyaient* 'believed, thought', the task is to correctly identify its French lemma *croire* 'believe, think'.

Yarowsky *et al.*'s morphology projection method approaches this problem by first examining the word alignments for this inflected French word *croyaient.* It is found that *croyaient* is aligned to the English words *believed* and *thought,* among others:

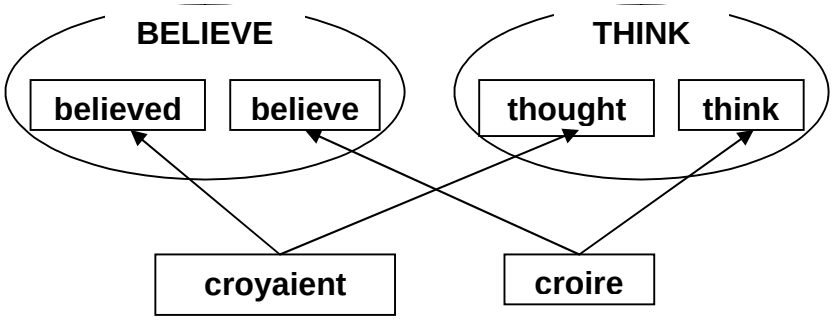


The alignments for each of these English words are then examined. Ideally, an alignment involving the lemma *croire* would be among them, and thus there would be a link from *croyaient* to *croire* via these word alignments with the shared English bridge words: for example, *croyaient* → *believed* → *croire.* Unfortunately, it is rare for inflected forms such as *believed* and *thought* to be aligned to lemmas such as *croire.* More commonly, lemmas are aligned with lemmas, and inflected forms are aligned with inflected forms.



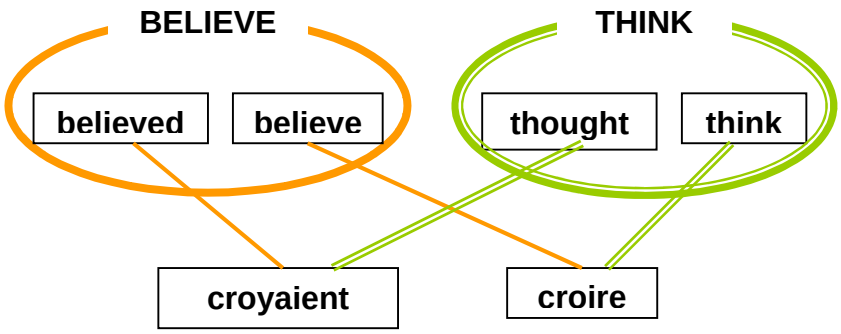
As shown above, the inflected French word *croyaient* is involved in alignments with the inflected English words *believed* and *thought*, while the French lemma *croire* is involved in alignments with the English lemmas *believe* and *think*. Crucially, there are no alignments between *croyaient* and *believe* and *think*, and no alignments between *croire* and *believed* and *thought*. Thus at this point *croyaient* and *croire* do not share any aligned English words in common that can serve as a bridge for morphology projection.

To overcome this impasse, a lemmatizer for the bridge language is used in order to collapse the inflected forms into their respective lemmas. In the above example, this means that the English side of the parallel corpus is lemmatized, with the effect that all occurrences of inflected forms such as *believed* and *thought* are replaced with their respective lemmas, such as *BELIEVE* and *THINK*. The resulting situation can be visualized as follows:



Thus lemmatizing reduces the number of unique English words – in the above example, the English vocabulary has been reduced to two lemmas *believe* and *think*. In terms of alignments, this means that the previous alignments (*croyaient, believed*), (*croyaient, thought*), (*croire, believe*), and (*croire, think*) now involve the English lemmas *BELIEVE* and *THINK*: (*croyaient, BELIEVE*), (*croyaient, THINK*), (*croire, BELIEVE*), and (*croire, THINK*). As these new alignments show, *croyaient* and *croire* now have aligned English words in common – namely, *BELIEVE* and *THINK*.

These common aligned English words now can serve as bridges linking *croyaient* to *croire*. As the diagram below shows, these two common aligned English words provide two paths linking *croyaient* and *croire*:



Thus, the morphological analysis applied to the English side of the corpus has been projected onto the French side via these lemmatized bridge words. The end result is that the French side of the corpus is now morphologically analyzed as well, with inflections such as *croyaient* linked to their respective lemmas, such as *croire*.

As mentioned earlier in section 1.2.2, this morphological projection is formalized in Yarowsky *et al.* (2001) as the following similarity measure:

$$P_{mp}(F_{root} \mid F_{infl}) = \Sigma_i P_a(F_{root} \mid E_{lem\,i})\, P_a(E_{lem\,i} \mid F_{infl})$$

where:

- P_{mp} refers to morphology projection probability. It expresses how probable a certain candidate French root (lemma) is for a given inflected French word, based on the similarity calculation involving their alignment probabilities.

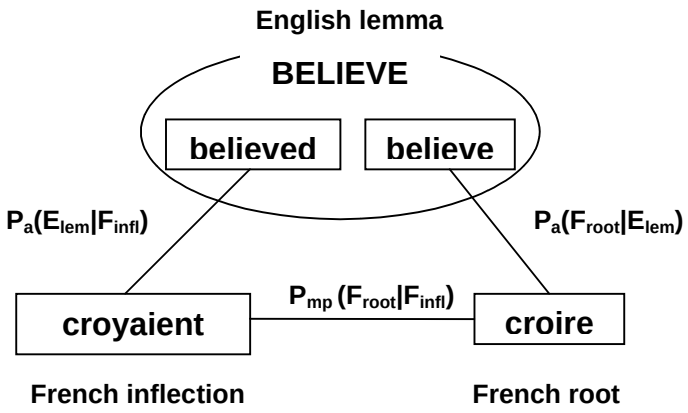
- P_a refers to alignment probability.

- F_{root} refers to candidate French root (lemma).

- F_{infl} refers to inflected French word.

- E_{lem} refers to English lemmas with which both the candidate French root and the inflected French word are aligned (i.e., the aligned English lemmas they have in common.). The index *i* indicates that the product of the alignment probabilities are summed over all such English lemmas which both the candidate French root and inflected French word share an alignment with. In other words, this calculation is performed for every English lemma that is involved in alignments with both the candidate French root and the inflected French word.

These probabilities can be visualized as follows:



Each of the alignment probabilities P_a can be broken down into the following calculations involving word alignment frequencies:

- $P_a(F_{root}|E_{lem}) = f(F_{root}, E_{lem}) \div f(E_{lem})$
- $P_a(E_{lem}|F_{infl}) = f(E_{lem}, F_{infl}) \div f(F_{infl})$

Thus the alignment probability between a French root and an English lemma, $P_a(F_{root}|E_{lem})$, is calculated by dividing the number of times this French root and English lemma are aligned together ($f(F_{root}, E_{lem})$) by the number of times this English lemma occurs in the corpus ($f(E_{lem})$).

Similarly, the alignment probability between an English lemma and a French inflection is calculated by dividing the number of times this English lemma and French inflection are aligned together ($f(E_{lem}, F_{infl})$) by the number of times this French inflection occurs in the corpus ($f(F_{infl})$).

In terms of word alignment frequencies, the overall morphology projection probability calculation can therefore be expressed as follows:

$$P_{mp}(F_{root}|F_{infl}) = \Sigma_i P_a(F_{root}|E_{lem\,i}) P_a(E_{lem\,i}|F_{infl})$$

$$= \Sigma_i \left(f(F_{root}, E_{lem}) \div f(E_{lem})\right) \times \left(f(E_{lem}, F_{infl}) \div f(F_{infl})\right)$$

The implementation of this method involved the following steps (this explanation assumes that the English side of the parallel corpus has already been lemmatized):

1. Processed the three separate files of the parallel corpus (the French text file, the English text file, and the word alignments file) so that aligned words were paired up together.

2. Counted the frequency of each unique word alignment.

3. Each unique French word *x* in the French side of the parallel corpus was treated as a French inflection for which the correct lemma had to be found. This word *x* was therefore systematically considered with every unique French word *y* in this same corpus. Each word *y* was treated a candidate lemma for the inflection *x*.

   For each pair of inflected French word and candidate French lemma <*x*,*y*>:

   i) Found the English lemmas with which both *x* and *y* were aligned (i.e. the English lemmas they had in common among their alignments).

   ii) For each of these shared English lemmas, collected the alignments involving *x* and the lemma, and *y* and the lemma – in other words, the alignments (*x*, English lemma) and (*y*, English lemma).

   iii) For each of these shared English lemmas, calculated the product $P_a(F_{root} | E_{lem})\ P_a(E_{lem} | F_{infl})$ using the relevant alignments collected

in step (ii) above. Word *x* is the F_{infl}, word *y* is the F_{root}, and E_{lem} is the shared English lemma. Thus, for each shared English lemma, the alignment frequencies of the relevant alignments (*x*, English lemma) and (*y*, English lemma) were used to calculate a product for this particular combination of *x, y,* and English lemma.

4. Summed all the products for each pair of inflected French word and candidate French lemma <*x*,*y*>. At the end of this step, each pair <*x*,*y*> was assigned a single sum value. This sum value was the morphology projection probability $P_{mp}(F_{root} | F_{infl})$ for this particular pair of F_{infl} and F_{root} *x* and *y*.

5. For each F_{infl} *x*, found the F_{root} *y* with which it had the highest projection probability value. This "best" F_{root} *y* was then considered as the French lemma determined by the Projection Method for the inflected French word *x*.

6. Collected all the inflected French words that were determined to belong to the same French lemma into its own conflation set. These sets of related words are the inflectional paradigms determined by the Projection Method.

Below is a sample of the conflation sets (one set per line) outputted by the Projection Method:

```
d\xe9crire
attend attendre
devions doit
parle
doit doive doivent m\xe9riterait
europ\xe9enne
assurant garantir
palestinienne palestiniens
a ai auraient aurais aurait aurions aurons avaient ava
eussent n\xe9cessiter ont parlez
frapp\xe9s
d\xe9pensant d\xe9penses
niveau niveaux prennent
parlions
assur\xe9e effectuer
fond\xe9e fonde
rechercher
```

Figure 6. Output of Projection method

Chapter 3

RESULTS

As the gold standard is comprised of 100 unique inflectional paradigms broken up into word pairs, evaluation was carried out by first taking the conflation sets outputted by each of the morphology methods and converting them into word pairs. These word pairs were then compared against those in the gold standard, and the number of correct word pairs calculated. This number was then used to calculate precision, recall, and f-measure scores. Each of these scores was calculated as follows:

- Precision = (# of Correct Word Pairs) ÷ (# of Word Pairs Evaluated)
  The precision score expresses how many of the method's outputted word pairs are correct (agrees with the gold standard).

- Recall = (# of Correct Word Pairs) ÷ (# of Word Pairs in Gold Standard)
  The recall score expresses how many of the word pairs in the gold standard the method is able to output (in other words, how well the method is able to output the amount of word pairs it is expected to be able to output, based on the total number of word pairs in the gold standard).

- F = (2 × Precision × Recall) ÷ (Precision + Recall)
  The f-measure is the harmonic mean of precision and recall, and it favors equal scores for both precision and recall. In other words, a high f-measure means that both precision and recall are high as well.

The evaluation of the morphology methods is quantified via these three scores.

3.1 Evaluation of the Prefix Similarity Method

As described in section 2.3.1, the Prefix Similarity method compares the prefixes of words using three rules, with the execution of each rule dependent upon the lengths of the two words being compared. The main rule is that if both words are four or more letters in length, if their first four letters are the same, then they are morphologically related. For words less than four letters in length, however, there are several rules and rule combinations that can be used:

Option a) (This is the best-performing rule combination):

If both words are either of length 1 or length 2, then if their first letters are the same, they are morphologically related. (e.g. [*a*, *a*], [*l*e, *l*a])

If both words are of length 3, then if their first two letters are the same, they are morphologically related. (e.g. [*di*s, *di*t])

Option b) (New rule for words of length 2; other rules same as in (a) above):

If both words are of length 1, and if they are exactly identical, then they are morphologically related. (e.g. [*a*, *a*])

If both words are of length 2, and if they are exactly identical, then they are morphologically related. (e.g. [*le*, *le*])

If both words are of length 3, then if their first two letters are the same, they are morphologically related. (e.g. [*di*s, *di*t])

Option c) (New rule for words of length 3; other rules same as in (a) above):

If both words are either of length 1 or length 2, then if their first letters are the same, they are morphologically related. (e.g. [*a*, *a*], [*l*e, *l*a])

If both words are of length 3, and if they are exactly identical, then they are morphologically related. (e.g. [*dis*, *dis*])

Option d) (Uses both new rules given above for words of length 2 and 3):

If both words are of length 1, and if they are exactly identical, then they

are morphologically related. (e.g. [*a*, *a*])

If both words are of length 2, and if they are exactly identical, then they are morphologically related. (e.g. [*le*, *le*])

If both words are of length 3, and if they are exactly identical, then they are morphologically related. (e.g. [*dis*, *dis*])

These different options were then each tested via four separate experiments using a 120,000-word subset of the French side of the parallel corpus (5164 lines of monolingual French text).[19] The results obtained are shown below, with the best-performing rule combination (option (a) above) highlighted:

Rule Combination	*# of Unique Words Evaluated*	*# of Unique Paradigms Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
a	391	145	871	790	90.7	48.8	63.5
b	391	146	872	789	90.5	48.7	63.3
c	391	147	873	788	90.3	48.7	63.2
d	391	148	874	787	90.0	48.6	63.1

Table 6. Scores for the Prefix Similarity method

For reference, the gold standard contains 391 unique words – thus there are 391 evaluable unique French words. As the second column in the above table shows, all 391 evaluable French types are present in the output of the Prefix Similarity method. From these 391 evaluable types, around 870 unique word pairs were able to be made. The gold standard contains 1,619 word pairs.

Making the rules determining prefix similarity more restrictive (as in rule combinations (b), (c), and (d)) led to an increase in the number of unique paradigms posited by the method (shown in column 3 above). In other words, as the rules became more restrictive, morphologically related words no longer got

[19] Section 2.1.1 explains the justification for using a 120,000-word subset French corpus instead of the whole French side of the parallel corpus.

lumped together – rather, they got put into their own sets. Thus the correct inflectional paradigms got fragmented, resulting in both an increase in the number of possible word pairs and a decrease in the number of correct word pairs. This resulted in a decrease in both precision and recall.

As highlighted in Table 6, the best-performing rule combination was option (a), which involved the least restrictive rules for determining prefix similarity. Since this rule combination yielded the best results, it is this version of the Prefix Similarity method which was used in subsequent experiments that combined all three morphology learning methods together.

The Prefix Similarity method achieved a high precision score (90.7%). This means that the paradigms it outputted were mostly correct. Considering what prefix matching does – it posits that words with similar prefixes are morphologically related – one can see why this is so. Words that start with the same letter sequence usually *are* morphologically related (consider the various inflected forms of the verb *parler* 'to speak'[20]: *parle, parles, parlons, parlez, parlent,* etc.) Thus the inflectional paradigms determined by this method end up being populated by words that are in fact morphologically related.

The recall score for the Prefix Similarity method, however, is strikingly lower at 48.8%. The method was only able to output 871 word pairs, while the number of word pairs it is expected to output is 1,619 (the number of word pairs in the gold standard). Thus there is a good proportion of morphological variants which this method is not able to lump together with their related words – in other words, there is a considerable amount of paradigm fragmentation. Outputting the expected number of word pairs depends on having the correct paradigms in the first place, since the pairing script creates all possible unique word pairs from a

[20] The full inflectional paradigm of *parler* is given in section 1.1.2.

given paradigm. If a paradigm is not correct in the first place, then not all expected word pairs from this paradigm will be able to be created. As Table 6 above shows, the Prefix Similarity method posits 145 unique paradigms, as opposed to the 100 correct paradigms in the gold standard which is expected of it. This means that there are 45 spurious paradigms in the output, which bring forth incorrect word pairs – lowering precision. The drop in recall, for its part, can be attributed to the fragmentation of the correct paradigms – some word pairs are never able to be created because their component words have been put into separate paradigms.

Another source of error is the fact that this method only considers *prefixes*. Thus, it does well with words that are inflected via suffixation, since these words have a stem that stays constant while only their suffixes change. However, doing straight prefix-matching also causes the method to lump together morphologically unrelated words that happen to have the same prefix (for example, *international* 'international' and *interdit* 'forbidden').

In addition, the Prefix Similarity method, being an orthographically-based method, cannot handle irregular morphology. Thus irregular morphological variants (such as *eu,* which belongs to the paradigm of *avoir* 'to have') never get placed into their correct paradigms.

Lastly, the Prefix Similarity method is implicitly limited to comparing words that are mostly of the same length. In this implementation, words that are 4 or more letters long can be compared freely amongst each other, but words that are less than 4 letters long can only be compared with words of the exact same length (1-letter words with 1-letter words only, 2-letter words with 2-letter words only, and 3-letter words with 3-letter words only). However, this is an inherent limitation involved in comparing prefixes – it is necessary to delineate a cutoff for what can be considered to be a prefix. For example, *a* and *avoir* are part of the same

paradigm, but it is difficult to group them together via prefix similarity, since in order to do so it would be necessary to posit that all words that begin with *a* are related. Thus, prefix similarity comparisons are limited by the fact that words have to be of around the same length in order for their "prefixes" to be reasonably compared.

So, as a baseline method, Prefix Similarity has a high precision and a middling recall score.

3.2 Evaluation of the Levenshtein Distance Method

As described in section 2.3.2, the Levenshtein Distance method determines the morphological similarity of a pair of words by calculating their string edit distance and seeing if this distance is less than or equal to a specified cutoff value. Several cutoff values, ranging from 1 to 5, were tried out in experiments. The best-performing version of the method had a cutoff value of 2.

The results from these experiments are shown below. As with the Prefix Similarity method, the experiments for the Levenshtein Distance method were done on a 120,000-word subset of the French side of the parallel corpus (5164 lines of monolingual French text). The best-performing version of the method is highlighted in the table below:

Cutoff Value for Levenshtein Distance	*# of Unique Words Evaluated*	*# of Unique Paradigms Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
1	391	325	485	339	69.9	20.9	32.2
2	391	305	1600	838	52.4	51.8	52.1
3	391	310	6711	1466	21.8	90.5	35.2
4	391	362	21,571	1594	7.4	98.5	13.7
5	391	376	46,142	1597	3.5	98.6	6.7

Table 7. Scores for the Levenshtein Distance method

Again, for reference: the gold standard contains 1,619 word pairs built up from 100 unique French inflectional paradigms that all in all contain 391 unique French words.

The optimal cutoff value for the Levenshtein Distance was found to be 2. At this value, the highest f-measure was obtained, and precision and recall were both at around the same percentage, 50%. With a harsher cutoff value of 1, precision is high (69.9%) since words that vary by only one string edit most likely are morphologically related. Recall, conversely, is a low 20.9% since lots of morphological variants that vary by more than one string edit from the word under consideration are ruled out because their distances are greater than the cutoff value.

As the cutoff value is made more and more lax, from 3 going up to 5, precision declines rapidly to 3.5%, since the less restrictive cutoff value allows more words to be grouped together. Naturally, within these less restrictive groupings, there are words that are not morphologically related and so are incorrectly lumped together. These incorrect paradigms lead to incorrect word pairs that lower the precision score. Recall, however, goes up rapidly as the cutoff value gets more lax, eventually reaching 98.6%. This is because the less restrictive cutoff values for Levenshtein Distance allow for bigger and bigger paradigms to be built. These bigger paradigms in turn allow for the creation of a greater number of word pairs, which means the output is able to cover more of the expected word pairs in the gold standard.

Note that the Levenshtein Distance method achieves greater recall scores than the Prefix Similarity method (which had a top recall score of 48.8%). This is due to the difference in the way these methods determine morphological similarity. The Prefix Similarity method can only group together words that have similar prefixes; in contrast, the Levenshtein Distance method considers overall string

edit distance and is thus not limited to considering just the beginning character sequence of a word. Thus the Levenshtein Distance method is able to posit more matches and consequently build paradigms that have more members. This greater number of members per paradigm results in a greater number of word pairs being created overall. (In the experiments, the Prefix Similarity method created at most 148 unique paradigms and 874 word pairs; in contrast, the Levenshtein Distance method created at most 346 unique paradigms and 46,142 word pairs.) The greater number of word pairs leads to an increase in recall scores.

As highlighted in Table 7 above, the optimal cutoff value for the Levenshtein Distance is 2. At this value, precision and recall are on equal footing, and so the f-measure obtained is the highest. At values lower and higher than 2, the tradeoff between precision and recall is manifested as a striking discrepancy between the two scores, leading to a drop in f-measure.

Here are the best results for each of the two monolingual morphology learning methods, for comparison purposes:

Morphology Learning Method	*# of Unique Words Evaluated*	*# of Unique Paradigms Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
Prefix Similarity	391	145	871	790	90.7	48.8	63.5
Levenshtein Distance	391	305	1600	838	52.4	51.8	52.1

Table 8. Best scores for monolingual morphology learning methods

The Prefix Similarity method has a higher f-measure due to its high precision score. The Levenshtein Distance method has a slightly higher recall score than the Prefix Similarity method, but its lower precision score keeps its f-measure at 52.1%. Overall, though, the performance of these two methods can be

summarized as follows: the Prefix Similarity method is more precise but lacking in recall, while the Levenshtein Distance method is capable of almost full recall but must sacrifice precision in order to achieve this.

The precision, recall, and f-measure scores in Table 8 above represent the baseline results in this study. The performance of the Projection Method, which utilizes parallel corpora, will be gauged against this baseline.

3.3 Evaluation of the Projection Method

As described in section 2.3.3, lemmatizing the bridge language side of the parallel corpus (which was English in this study) leads to an increase in the performance of the Morphology Projection method. This is because lemmatizing converts inflections into their lemmas, thus increasing the chances that a given French inflection and a candidate French root will have an aligned English lemma in common. These shared aligned English lemmas serve as bridges for morphology projection, so the more shared aligned English lemmas there are for a given <inflection, candidate root> pair, the better the chances that the similarity calculation will be able to correctly identify whether this pair is in fact validly related or not.

The English lemmatizer used in this study was *morpha* (Minnen *et al.* (2001)), described in detail in section 2.1.2. Since *morpha* accepts both POS-tagged and untagged data as input for lemmatization, experiments were run to see what difference, if any, tagged input data makes with respect to the performance of the Projection method. Minnen *et al.* recommend the use of POS-tagged data as input to *morpha,* for maximum lemmatizing accuracy.

Three experiments are shown in the table below. These experiments involve a subset of the French-English parallel corpus – specifically, 5164 lines of French text and 5164 lines of English text, which is approximately 120,000 words for

each language. The first experiment involved no tagging or lemmatizing of the English data. For the last two experiments, the English data was lemmatized, but one experiment involved tagging while the other did not. The Projection method performed as follows with respect to these three different kinds of input data, with the best-performing version highlighted:

English Data Used	*# of Unique Words Evaluated*	*# of Unique Paradigms Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
Untagged Unlemmatized	218	192	235	62	26.4	3.8	6.7
Untagged Lemmatized	193	149	374	153	40.9	9.5	15.4
Tagged Lemmatized	191	148	373	149	39.9	9.2	15.0

Table 9. Scores for the Projection method

The first thing to note is that lemmatizing the bridge language definitely improves the performance of the Projection method. Using unlemmatized English data resulted in a recall score of only 3.8% and a precision score of 26.4%. With lemmatized English data, however, recall jumped to 9.2% and 9.5%, and precision to 39.9% and 40.9%. The number of spurious paradigms also decreases with lemmatization, from 192 for unlemmatized data, to 148 and 149 for lemmatized data (the gold standard has 100 paradigms). This indicates a decrease in paradigm fragmentation – with lemmatization, words are correctly being grouped into their proper paradigms and not into spurious paradigms all by themselves.

The second thing to note is that tagging the input data to the lemmatizer actually resulted in a slight drop in scores – from 40.9% precision and 9.5% recall for untagged data, to 39.9% precision and 9.2% recall for tagged data. However, overall the difference in performance between using tagged and untagged English

data is quite small – only a one-point difference in precision, 0.3-point difference in recall, and 0.4-point difference in f-measure. Thus, tagging the English data did not bring about any improvement in performance and actually harmed it a little. Because of this, it is preferable to use untagged bridge language data for the Projection method, as such data was found to bring about slightly higher performance scores. Furthermore, using untagged data is in keeping with the unsupervised nature of this study's approach to morphology learning – eliminating the need for a POS-tagger means using one less knowledge source. Since some bridge languages might not have existing POS-taggers, this is a desirable situation to maintain.

The best scores for each of the three morphology learning methods are shown in the table below:

Morphology Learning Method	*# of Unique Words Evaluated*	*# of Unique Paradigms Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
Prefix Similarity	391	145	871	790	90.7	48.8	63.5
Levenshtein Distance	391	305	1600	838	52.4	51.8	52.1
Morphology Projection	193	149	374	153	40.9	9.5	15.4

Table 10. Best scores for all three methods

As the table above shows, the Projection method achieves the lowest precision and recall scores out of the three methods. Again, for reference: the gold standard contains 1,619 word pairs built up from 100 unique French inflectional paradigms that all in all contain 391 unique French words. Both of the baseline, monolingual morphology learning methods (Prefix Similarity and Levenshtein Distance) have all 391 evaluable French types in their output. The Projection method, however, only has 193 of these evaluable types in its evaluated output.

Consequently, the evaluated paradigms from the Projection method do not contain all the words they are supposed to contain. The method outputs 149 unique inflectional paradigms (column 3 in the table above), which is around the same number outputted by the Prefix Similarity method. However, because the Projection method's output contains a smaller number of unique French words than the Prefix Similarity method's output, the paradigms produced by the Projection method contain fewer members. This is discernible in the number of word pairs created by each method: even though they have roughly the same number of unique paradigms, Projection is able to create only 374 word pairs from these paradigms, while Prefix Similarity creates 871.

Thus, the failure of the Projection method to output the full range of evaluable unique French words gives rise to a low number of word pairs being created, which in turn results in a dismal recall score (9.5%).

So where do all the unique French words go? How come they don't show up in the output of the Projection method? The low number of unique French words being outputted by the Projection method can be attributed to three things: 1) the inherent requirement that a given French inflection and its candidate root must share at least one aligned English lemma in common for the Projection method to even calculate their morphological similarity, 2) weak statistics, and 3) the inherent requirement that both words in a word pair must be words present in the gold standard in order for the word pair to be evaluable.

In order for the Projection method to do the morphology similarity calculation for a given <inflected French word, candidate French root> pair, these two French words must have at least one aligned English lemma in common. The probability calculation depends on this, since this is how it selects which alignments to consider for the frequency calculations. If the inflected French word and candidate French root do not have an aligned English lemma in

common, then the Projection method cannot do the similarity calculation for this word pair and thus cannot ever posit a relation between these two words.

This is what happens with the word *carence* 'deficiency' and its plural form *carences*. Both words occur in the gold standard, but do not occur in the output of the Projection method. This is why. The alignments these words are involved in are as follows (given in the format French Word, English Word, Alignment Frequency):

- Alignments *carence* is involved in:
  `carence unachieve 1`
  `carence task 1`

- Alignments *carences* is involved in:
  `carences lot 1`
  `carences grind 1`

As the above shows, *carence* and *carences* do not have an aligned English lemma in common. This means that the Projection method will never be able to perform the similarity calculation between *carence* and *carences*. Thus there is already no chance of *carences* ever being linked to its correct French root *carence*.

Furthermore, *carence* and *carences* don't even get linked to themselves as their best possible roots. Because of weak statistics, incorrect alignments with high frequencies of occurrence get chosen instead as the best roots for these words. For example, the probability for *carence* being the root word for itself is 0.03, but the probability for *tâche* 'task' being the root for *carence* is 0.13, which is greater. Similarly, the probability for *carences* being the root word for itself is 0.13, but the probability for *beaucoup* 'much' being the root word for *carences* is greater at 0.33. Thus, *tâche* is chosen as the root word for *carence* and *beaucoup* as the root word for *carences*.

This is bad news come evaluation time, because the words *tâche* and *beaucoup* are not in the gold standard. Thus, the word pairs (*carence, tâche*) and (*carences, beaucoup*) are not evaluable. Before evaluation is performed, all word pairs in the output are scrutinized, and those which contain words not in the gold standard are excluded from consideration. This is why *carence* and *carences* never show up in the evaluated word pairs

Thus, it is not a problem of *outputting* all 391 French types in the gold standard – the Projection method does do that, and does find the best candidate roots for each of these 391 words. The problem is that these types are paired with candidate roots that are not in the gold standard and are thus not evaluable. Thus the word pairs that they are a part of are in turn also not evaluable. So it is a consequence of the evaluation strategy – maintaining an evaluable vs. non-evaluable distinction – that such a large proportion of French types are excluded from the evaluation (even though by themselves they *are* present in the gold standard and are thus evaluable in themselves). Using a gold standard means that evaluation is always limited to a select portion of the data. In the case of the Projection method, this translates to a loss of French types.

Mostly, however, weak statistics is to blame for the poor performance of the Projection method. Since the similarity calculation is dependent upon alignment frequency, a given French inflection commonly gets linked to a French root that happens to participate in a high-frequency alignment with one of the aligned English lemmas that the inflection and the root share. This is how *carences* got linked to *beaucoup* – out of all the other candidate French roots which, with *carences,* shared an alignment with the English lemma *LOT*, *beaucoup* had the highest frequency – 6, in comparison to the other roots, which all had an alignment frequency of 1 with *LOT*. Thus *beaucoup* ended up with the highest probability value after the similarity calculation was performed.

Also because of weak statistics, there are instances in which the best-scoring candidate root for a given word is the word itself. For example, *danger* 'danger' is found to have itself as its best root, while its plural form *dangers* is similarly found to have itself (*dangers*) as its best root. Consequently, these words get placed in separate paradigms: *danger* is put in its own paradigm [*danger*], while *dangers* is similarly all by itself in its own paradigm [*dangers*]. Thus the correct word pair (*danger, dangers*) never gets formed, since these words are in separate paradigms. Such paradigm fragmentation due to weak statistics is a main reason for the low recall score achieved by the Projection method.

As described in section 2.1.1, the word alignments for the parallel corpus were automatically created using GIZA++ (Och and Ney (2003)). Because of this, there are incorrect alignments, most of which have a frequency of 1. A separate experiment was done to see if weeding out these low-frequency alignments would lead to any improvement in the performance of the Projection method. The corpus used in this experiment was the same as with the earlier Projection method experiments (5164 lines each of French and English, about 120,000 words per language), and the type of English data used was the one that was found to yield the best results (untagged and lemmatized). Alignments with a frequency count of 1 were excluded from consideration. The result of the experiment is shown below:

Method	*# of Unique Words Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
Projection with Alignment Weeding	91	203	137	67.5	8.5	15.0

Table 11. Scores for Projection method with alignment weeding

Weeding out low-frequency alignments in this way resulted in an improved precision score (from 40.9% when no weeding is done, to 67.5%). This is because false matches are eliminated before the calculations are done, thus reducing the number of incorrect candidate roots. However, this increased precision is achieved at the expense of recall (which drops from 9.5% when no weeding is done, to 8.5%). There are fewer French types evaluated (from 193 when no weeding is done, to 91) and consequently fewer word pairs evaluated (from 374 to 203). This is because the weeding, while getting rid of incorrect alignments, also gets rid of some valid ones that happen to have a frequency count of 1. Thus, some links are never made for certain <inflection, root> pairs because the alignments involving an English lemma that they have in common happened to be weeded away.

What the Projection method does do well, however, is learn irregular morphology. The baseline, monolingual morphology learning methods (Prefix Similarity and Levenshtein Distance) are unable to handle irregular morphology because they rely on orthography to determine morphological relatedness. In contrast, the Projection method is not at all dependent on orthography and instead uses word alignment frequencies to judge the morphological relatedness of words. While this reliance on statistics can be a weakness (as the preceding paragraphs discuss), with respect to the learning of irregular morphology it is the Projection method's singular strength.

For example, the Projection method is able to correctly group together these inflected forms of the irregular verb *avoir* 'to have':

```
a ai auraient aurais aurait aurions aurons avaient
avais avons ayons eus eussent ont
```

The best-performing version of the Projection method is the one with scores given in Table 10 – the one using untagged, lemmatized English data with no

weeding of low-frequency alignments. The method suffers from poor recall due to weak statistics and the lack of a shared aligned English lemma between a French inflection and its candidate French root. Precision can be increased with weeding of low-frequency alignments, but this leads to an undesirable further drop in recall. Thus, to improve the Projection method one can try either 1) using more data to overcome the effects of weak statistics, or 2) augmenting the Projection method with another morphology learning method capable of increasing precision, recall, or both.

3.4 Combining the methods together

Since the monolingual morphology learning methods achieved higher precision and recall scores than the Projection method, these methods were combined with the Projection method to see if any improvement in performance would result. The combination was done simply by running each of the methods and then pooling their outputted word pairs together for evaluation. The corpus involved was the same subset used in the previously described experiments. Only the best-performing version of each method was tested. The results were as follows:

Method	*# of Unique Words Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
Combined	391	2185	1148	52.5	70.9	60.4

Table 12. Scores for combined methods

As the above table shows, combining the methods resulted in a dramatic increase in recall (from 9.5% for the Projection method, to 70.9% for the Combined method). This is due to the fact that the word pairs outputted by the three methods were pooled together. Since the monolingual morphology learning methods, especially the Levenshtein Distance method, are capable of achieving recall scores of around 50%, the Combined method can therefore be expected to have a recall score of at least this percentage. Note also that because the output

of the three methods is combined in this way, all 391 gold standard French types are present in the evaluated data.

Precision, however, does not increase as much (from 40.9% for the Projection method, to 52.5% for the Combined method). This, again, is due to the way the combination of the methods was done. Pooling the outputted word pairs together increased recall because it brought forth more word pairs to be evaluated; however, this simple pooling means that incorrect word pairs are preserved as well. Because these incorrect word pairs are carried forth as-is into the final evaluation pool, the mistakes made by each of the three methods continue to hurt the precision of the Combined method. Combining the methods in a more sophisticated way (for example, ranking the best candidate root volunteered by each method and scaling these ranks to arrive at a combined score for that candidate root, as in Yarowsky and Wicentowski (2000)) would probably yield improved results.

3.5 Increasing the corpus size

Yarowsky *et al.* (2001) report that the Projection method achieves greater coverage when the corpus size is increased. To investigate this, an experiment was run using a corpus ten times bigger than the subset used in the previously described experiments. This subset was comprised of 51640 lines of English text and 51640 lines of French text – approximately 1 million words per language. The results for this experiment are shown below:

Method	*# of Unique Words Evaluated*	*# of Word Pairs Evaluated*	*# of Correct Word Pairs*	*Precision*	*Recall*	*F*
Prefix Similarity	391	871	790	90.7	48.8	63.5
Levenshtein Distance	391	1600	838	52.4	51.8	52.1
Morphology	194	286	149	52.1	9.2	15.6

Projection						
Combined	391	2105	1138	54.1	70.3	61.1

Table 13. Scores for increased corpus size

Increasing the corpus size brought about an increase in precision for the Projection method (from 40.9% using the smaller corpus, to 52.1%). While the number of French types in the evaluated output remained relatively unchanged (193 for the smaller corpus, 194 for the bigger corpus), the number of word pairs created was significantly less when the bigger corpus was used (286, as opposed to 374 for the smaller corpus). This indicates that there is a lesser degree of paradigm fragmentation when a bigger corpus is used. The fewer number of word pairs that result from the lessened number of spurious paradigms contributes to the increase in precision.

However, the fewer number of word pairs leads to a slight decrease in recall as well (from 9.5% for the smaller corpus, to 9.2%). Overall, though, f-measure goes up from 15.4% to 15.6%.

The performance of the monolingual morphology learning methods (Prefix Similarity and Levenshtein Distance) was not affected by the increase in corpus size – the scores achieved by these methods were exactly identical to the scores they obtained using the smaller corpus.[21] This is because with the smaller corpus, these methods have already "seen" all 391 unique French words in the gold standard, and grouped these words accordingly into paradigms. Thus the addition of more data has no further effect, since these orthography-based methods have already "seen" all the evaluable French types and segregated them accordingly into paradigms. Since they do not use statistics at all but are instead orthography-based, their evaluable performance with respect to the 391 French

[21] These scores are displayed in Table 8.

types in the gold standard derives no further benefit from the addition of more data (unlike the Projection method).

As for the effect of the increased corpus size on the Combined method, Table 13 shows that there is a slight increase in precision (from 52.5% for the Combined method using the smaller corpus, to 54.1%), a slight drop in recall (from 70.9% to 70.3%) and an overall increase in f-measure (from 60.4% to 61.1%). These changes mirror the changes seen in the scores for the Precision method, which was the only method affected by the increase in corpus size.

Chapter 4

CONCLUSION

4.1 Main observations

This study investigated the use of parallel corpora in learning inflectional French morphology. Three separate unsupervised morphology induction methods were implemented: the Prefix Similarity method, the Levenshtein Distance method, and the Morphology Projection method. The Prefix Similarity and Levenshtein Distance methods made use of only monolingual corpora and based their determinations of morphological relatedness on orthographical similarity. In contrast, the Projection method made use of parallel corpora and the word alignments contained therein to calculate the morphological relatedness of words.

Each method had its weaknesses and strengths. The monolingual methods, being orthographically-based, were capable of high precision and recall scores. This is because most words that are orthographically similar are also morphologically related. However, reliance on orthography also made these methods completely incapable of learning irregular morphology.

The Projection method, which was the only method to use parallel corpora, proved to be susceptible to the effects of weak statistics. It was also limited by the nature of the similarity calculation – the inherent requirement that a given French inflection and its candidate root must have an aligned English lemma in common. Many French inflections were not paired up with their correct roots because they simply did not have an aligned English lemma in common that could act as a bridge for morphology projection. Thus, it is best to supplement the Projection method with two things: 1) more data (increases in corpus size),

and 2) other morphology learning methods which can cover the gaps in the Projection method's output brought about by weak statistics.

Despite its weaknesses, the Projection method does do a good job of learning irregular morphology. In this respect, its reliance on statistics and complete disregard for orthography is a strength. This is another reason why it would be ideal to combine the Projection method with another method, perhaps one that is orthography-based – doing so would enable their strengths to cover each other's weaknesses.

4.2 Suggestions for improvement

Since model combination seems to be the best way to improve performance, it would be desirable to devise a more sophisticated way of combining the methods together, instead of the simple output pooling employed in this study. The candidate roots outputted by each method for a given French inflection can be ranked, the models itself weighted, and the scores scaled so that a final score for each root can be arrived at (*à la* Yarowsky and Wicentowski (2000)).

In addition, other morphology learning methods should be investigated, particularly those which make use of syntactic and semantic cues. Such methods are like the Projection method in that they are not orthography-based; however, they differ from the Projection method in what they take into account when determining morphological similarity. Syntactic context and semantic relationships have been shown to be good indicators of morphological relatedness (Yarowsky and Wicentowski (2000); Schone and Jurafsky (2001)). Methods that make use of such cues would probably contribute to improved performance, since they rely on features other than word alignment frequency.

Lastly, other evaluation strategies could be used to more closely examine the performance of the methods with respect to a certain part of speech class. For

example, Yarowsky *et al.* (2001) conducted their evaluation by looking only at verbs. Because of this, the results obtained in this study for the Projection method are not directly comparable with theirs, since the evaluation strategy is different. Evaluating with a more narrow perspective would enable these results to be more directly compared with those obtained in previous studies, and might also shed light on which kinds of inflections the methods do well with and which they have problems with.

4.3 Future work

In this study, the use of parallel corpora was found to facilitate the learning of irregular morphology. An additional thing to investigate would be the effect of using multiple parallel corpora for morphology projection. Yarowsky *et al.* (2001) did this, first using a French-English parallel corpus to project morphological information from English to French. Then, once the morphological analysis of the French was complete, it was used as a source of additional bridges for morphology projection into Spanish. It was found that the use of parallel corpora contributed to improved performance of their morphology learning method.

Since the Projection method in this study could still benefit from additional improvements, this step (using multiple parallel corpora) could not be attempted since the French was not yet satisfactorily morphologically analyzed. Once the additional methods and improvements suggested above are implemented, however, hopefully the performance of the Projection method will improve and the French will be sufficiently morphologically analyzed for it to be used as a bridge language for morphology projection.

BIBLIOGRAPHY

Baayen, R. H., R. Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database (CD-ROM). LDC, University of Pennsylvania.

Callison-Burch, C. and C. Bannard. 2005. Paraphrasing with bilingual parallel corpora. In Proceedings of ACL-2005.

Jurafsky, D. and J. Martin. 2000. *Speech and Language Processing.* Prentice-Hall.

Kendris, C. 1996. (4th ed.) *501 French Verbs.* Barron's Educational Series, Inc.

Koehn, P. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished draft.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit 2005.

Lessard, G. 1996. *FREN 215: Introduction à la linguistique française.* WWW document. URL http://post.queensu.ca/~lessardg/Cours/215/ Accessed 16 August 2006.

Minnen, G., J. Carroll, and D. Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering.* 7/3: 207-223.

Och, F. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics.* 29/1: 19-51.

Rogati, M., S. McCarley, and Y. Yang. 2003. Unsupervised learning of Arabic stemming using a parallel corpus. In Proceedings of ACL-2003.

Romary, L., S. Salmon-Alt, and G. Francopoulo. 2004. Standards going concrete: from LMF to Morphalou. COLING-2004.

Schone, P. and D. Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In Proceedings of NAACL-2001.

Yarowsky, D., G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In Proceedings of HLT-2001.

Yarowsky, D. and R. Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In Proceedings of ACL-2000.